SMU Data Science Review

Volume 2 | Number 3 Article 8

2019

Analyzing Influences on U.S. Baby Name Trends

Laura Ludwig
Southern Methodist University, lludwig@smu.edu

Mallory Hightower
Southern Methodist University, mhightower@smu.edu

Daniel W. Engels
Southern Methodist University, dwe@smu.edu

Monnie McGee
Southern Methodist University, mmcgee@smu.edu

Analyzing Influences on U.S. Baby Name Trends

Mallory Hightower, Laura Ludwig, Daniel W. Engels, and Monnie McGee

Southern Methodist University, Dallas TX 75205, USA
{mhightower, lludwig, dwe, mmcgee}@smu.edu

Abstract. In this paper, an analysis is presented of the changing baby namespace and a model is created for predicting if a name's popularity is trending up or down. Just as cultures and societies change over time, baby names evolve to reflect these changes. By analyzing name phonemes and historical influences, one can better understand the underlying causes of the changing name trend. Utilizing the U.S. Social Security Administration (SSA) name registry and historical figure data sets, the influence of historical figures and name pronunciation on the naming trend was examined. Two neural networks were created to predict name trend, one utilizing name count and the other utilizing name pronunciation. Phoneme embeddings were also created to cluster and visualize similar and dissimilar sounding names. The analysis concluded that while historical factors do influence the U.S. naming trend, these factors are too inconsistent and sporadic to include in a name forecasting model. The phoneme-driven model classified name trend with 72% percent accuracy, while the model using name counts achieved 92% accuracy. Based on these results, there is a relationship between similar sounding names and their popularity trends, but it is not as predictive as purely using name count.

1 Introduction

Throughout history, there has been an interest in the causes and underlying motivations of the changes in the namespace. Historians, scientists, and statisticians alike have tried to answer the famous question posed by Shakespeare in 1597, *"What's in a Name?"*[1]. Throughout history the commonality of names has greatly changed. Names often shift in popularity over time and between genders, sometimes even disappearing completely from the namespace. Names have the ability to carry unique historical significance, potentially holding hundreds of years of history and information within them[2]. By understanding the driving forces behind the changing namespace, one can better predict name popularity trends.

[1] More information about Romeo and Juliet may be found at https://www.shakespeare.org.uk/explore-shakespeare/shakespedia/shakespeares-plays/romeo-and-juliet/. Last accessed 13 July 2019.

[2] For an example of how this topic is addressed socially, visit https://heleo.com/tim-urban-surprising-reasons-baby-names-become-trendy/5705/.Last accessed 13 July 2019.

Many potential reasons exist that could explain the shift in name popularity over time. Certain historical events such as natural disasters, the rise in pop culture icons, and new technology[3] all potentially play a role in influencing children's names. Other global developments such as immigration and the mixing of cultures also have the potential to alter the popularity of a name. Conversely, some factors seem to act as stabilizing influences on the popularity of a name, such as religion.

There has been consistent interest in studying the changing namespace, but most of this research focuses on understanding the relationship between names and topics such as race [6], economic advancement [16], and cultural events [8], without the practical application of these research findings. There are additional studies that consider the individual sound units of a name as a potential driving force behind the name's popularity [8]. However, the relationship between historical figures, the sound components of names, and the popularity of names, particularly concerning the useful application of the gained insight, remains under-analyzed. Examining the combination of these factors provides the ability to generate a robust prediction of name popularity to assist prospective parents in making a name decision that aligns with their priorities.

In order to create a useful application for the underlying factors that influence the naming trend, U.S. baby names, historical figures and events, and name phonemes were analyzed. The relationships between historical figures and the popularity of a name given to a child were visualized. The relationship between a name's phoneme set, the individual sound components of a name, and the resulting popularity of that name based on similar sounding names was also analyzed. The result was compared against a baseline model that utilized name count rather than sounds, and the embedding layer weights of the phoneme model were visualized to cluster similar sounding names.

The sources of this analysis are U.S. baby names and popular historical and cultural figures. The U.S. baby names were obtained from the U.S. Social Security Administration (SSA), with recorded baby names spanning from 1880-2018. The SSA data provides an extensive time series data set as it spans over a century. The historical figures data set was compiled from multiple online sources. Famous authors and pop culture characters, famous actors and actresses, famous athletes, and Biblical figures were included.

The historical figures data set was used as part of the historical name trend analysis. Name trends were visualized in the context of the presence of historical figures to determine if there was any substantial or long-term influence on the naming trend.

The name phonemes were gathered from a pronouncing dictionary. This created a data set of name and phoneme vectors, which were fed into the model for predicting a name's popularity based on the popularity of other similar sounding names.

[3] An example of using neural networks to generate new names can be found here: https://medium.com/@nateparrott/give-your-kids-futuristic-names-with-a-neural-network-9078bed0894d. Last accessed 13 July 2019.

After examining and visualizing the influence of multiple historical figures on the naming trend, the analysis concluded that there is no long-term or significant influence from these historical figures. The influence, when present, is small in volume and random to the extent that it cannot be quantified in a meaningful way that assists in predicting name popularity. There are some interesting, isolated influencers: a spike in the girl name Diana shortly after Lady Diana started dating Prince Charles in 1980, and the decrease in baby boys with the name Osama after Osama Bin Laden was first indicted by the U.S. Supreme Court. Overall, these groupings of historical figures do not appear to have significant influence on the naming trend.

Using name phonemes to predict a name's popularity proved promising but inconclusive. However, this model serves as a useful model with which to compare the baseline, count-driven model. Both models were trained with 20 epochs. The names only model utilized years 1951 to 2018 in the SSA data to predict the name trend, while the phoneme model only used the most recent 13 years (2006-2018). The count-driven model achieved a final accuracy of 92% and loss of 19%, and the phoneme model achieved a final accuracy of 72% and loss of 54%.

The goal of this analysis was to find significant factors for predicting the popularity trend of a name and to visualize the U.S. baby name trend at large. While the influence of various historical and popular figures proved non-substantial and sporadic at best, visualizations showed a few surprising results. Popular people in history do affect the U.S. naming trend, but not in a significant or meaningful enough way to be included in a model for predicting name popularity.

The phoneme LSTM model demonstrates that a name's pronunciation has value for predicting the name's popularity trend. However, the value is questionable when compared to the model that purely uses name count. The strong performance of the phoneme model, despite the fact that it was only trained on 13 years of data, in comparison to the count-driven model, which was trained on 68 years of data, concludes that there is an alternate and effective way to predict name popularity other than the traditional method of simply using the number of times that a name appeared in the past. However, the phoneme model may not be the superior model, but this is uncertain as computational constraints prevented from training the model on more years of data. While the relationship between the popularity of a name and the popularity of similar sounding names is not quantified in this analysis, there is definitely a strong relationship as demonstrated by the performance and visualization of the phoneme model.

The remainder of this paper is organized as follows. In Section 2 a discussion of related studies that analyze various influences on the naming trend is presented. In Section 3 a high level tutorial on neural networks and word embeddings and how they are utilized in this paper is presented. Discussion on the data sources and data attributes are discussed in Section 4. The approaches utilized in this analysis are discussed in Sections 5 and 6. Section 7 presents the overall results of the analyses and relevant visualizations. Discussion of the results and visualizations is presented in Section 8. The ethics of working with

name data are presented in Section 9 and relevant conclusions are drawn in Section 10.

2 Factors That Influence the Naming Trend

There has been extensive and conflicting research on the underlying factors and consequences of a child's name. Many of these conflicting study results concern the relationship between a baby's name and its race. In a study done to evaluate the effects of names that sound African-American, *The Causes And Consequences Of Distinctively Black Names* by Fryer and Levitt, the authors argue that people with traditional African-American names have no noticeable effects on their economic life [6]. However, an article in the *National Bureau of Economic Research* claims that it is indeed harder for a person with a name that sounds African-American to obtain employment than a person with a name that derives from Caucasian heritage [5]. The author of this article, David Francis, states that "Job applicants with white names needed to send about 10 resumes to get one callback; those with African-American names needed to send around 15 resumes to get one callback. ... [A] white name yields as many more callbacks as an additional eight years of experience." [5]. These works are just an example of the potential economic impacts produced by names. While these insights into the relationship between names, race, and economic success of a person are intriguing and important to understand as possible confounding variables, this analysis did not focus on the race of the names. The historical influence on names, rather than the racial influence, is the focus of the analysis.

There are multiple studies that demonstrate conflicting opinions on the influential power of a person's name to define their personality or their success in life. Stephen Dubner and Steven Levitt highlight a surprising example of defying the implications of a first name a in their *Freakonomics* chapter entitled "A Roshanda By Any Other Name" [10]. By contrast, Elisabeth Vincentelli articulates evidence of the opposite phenomenon in her *New York Times* article "You Are What Your Name Says You Are" [16]. Rather than becoming embroiled in taking a stance on this particular debate, this analysis focused on how contemporary and historical icons influence name trends as well as the actual phonetic pronunciation of the names.

Understanding and researching historical events is also necessary for this analysis. Such events have the potential to impact the naming of a child, as discussed in subsequent sections. Surveys are one way to gauge the opinion on the most impactful historical events. One such survey from Pew Research Center lists the ten most cited events that most profoundly affected America. Among the most impactful events are the September 11 attack, the election of Obama, the technology revolution, the fall of the Berlin Wall, and the moon landing. This article also points out that the historical events hold different weights for the unique subsections of the American demographic. For example, "the barrier-breaking Apollo 11 moon landing in 1969 ranks highest among Baby Boomers (35%), followed by 29% of those in the Silent and Greatest generations. The

momentous event does not register significantly with those in Generation X" [3]. This paper did not focus on the impact of historical events on the naming conventions of certain subgroups of Americans, but rather on the American population as a whole. Overall generational conclusions might be inferred based on the typical child-naming ages of parents, although the differences between generations in this analysis of name trends isn't discussed.

Names of past historical figures also have the potential to exert influence on naming trends. Kuakowskia et al., in their paper "Naming Boys after U.S. Presidents in 20th Century" discuss the influence of the U.S. presidents on the naming of American baby boys [9]. This paper creates an index formula to measure the influence of the various president names. The paper finds that "the fashion of naming babies after the actual American president passed away in 60s of XX century, while the fashion of naming babies after a favourite celebrity has remained" [9]. While the impact of U.S. presidents on the naming trend is not discussed here, the paper investigates the impact of actors and actresses, authors, pop culture figures, and other culturally and historically significant groups and events. The analysis [9] by Kuakowskia et al. is used as the basis of this analysis for creating an index function to measure historical influence on name trend.

Another significant group of characters in history that is frequently analyzed concerning its affect on the naming trend, are religious figures. The far reaching influence of the Catholic Church is demonstrated by the fact that "by the mid-1500s, religious names made up about half of all boys names" [2]. Even girls were given male saint names because there were fewer female saints in the Bible. In the 1500's, with the advent of the Puritans, people began giving their children "virtue names such as Grace, Faith, Hope, Charity" to "distinguish their children from what they saw as the godless masses" [2]. This paper also analyzes the affect of Biblical figures on the naming trend.

Additional research focuses on analyzing names by breaking the names down into their components, called phonemes. The phonemes are the individual sound components of a name, which when pronounced make one name unique from another name [14]. The way a name sounds also has also the potential to influence the naming trend. Burger et al. discuss the use of a name's phoneme to predict its popularity [8]. Among other things, the paper concluded that "names are more likely to become popular when similar names have been popular recently" [8]. The Burger at al. research [8] was used for inspiration to further investigate the similar or dissimilar traits between names. This paper continues the analysis by creating phoneme vectors and using those vectors to train a neural network. One of the byproducts of the neural network is a word embedding layer, which was used to analyze similar sounding name clusters.

3 Key Concepts for Understanding Neural Networks

3.1 Neural Networks

Neural networks are defined as "a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns" [11]. Two Long Short-

Term Memory (LSTM) networks are implemented, a form of the recurrent neural network, to predict the popularity trend of a name in this analysis. Unlike other recurrent neural networks, LSTMs do not suffer from the problem of vanishing gradients, where information early on in a long sequence of input is forgotten. This makes LSTMs a perfect candidate for time series analysis. While neural networks often perform extremely well at their machine learning tasks, they are often considered "black box" models because it is difficult to understand and interpret the model [18]. But because neural networks perform well in natural language processing tasks, time series tasks, and at approximating unknown functions, they are implemented in this analysis.

3.2 Word Embeddings

Word embeddings are "a type of word representation that allows words with similar meaning to have a similar representation" [1]. Word embeddings are especially useful for Natural Language Processing because they provide a method for transforming textual data into a format that a computer can process. One can then perform mathematical operations on the word embedding vectors, such as comparing the vectors with the Euclidean distance metric.

Phoneme vectors are created in this analysis to represent a name as a sequence of numbers that encodes the pronunciation. These phoneme vectors are then passed into the second LSTM model for mapping into a new embedding layer. The phonemes of the names are treated as if they were the words of a sentence. The neural network then embeds those inputs into weights that it can understand and use to determine the characteristics of a new name input.

The weights of the phoneme model embedding layer are essentially vector representations of the name pronunciation, as determined by the network in back propagation to minimize loss. This results in similar sounding names having similar vector representations. For example, Bob and Bobby would theoretically have very similar vector representations, while Bob and Timothy would not. Some word embedding models can even learn to distinguish gender, which some view as "disturbing" and a source of bias in a model [15]. However, for the purpose of mapping name embeddings, a model that distinguishes between gender may prove useful. By clustering the vectors from the phoneme model embedding layer on a 2D plot, the relationship between similar sounding names can be easily visualized.

4 Data

4.1 Social Security Administration Data

The data used in this analysis was acquired from the Social Security Administration (SSA) [7]. There are two types of data available from the SSA: yearly, national data files, and per-state files with yearly details. The national files, one per year, list the name, gender and count of people born in that year with a

given name. The per-state files list the name, the year of birth, and the count of people born that year with that name. Combining these files allowed us to create a data set suitable to conducting time series analyses.

The SSA data files were aggregated to create a comprehensive data set that contains the name of the baby (Name), the gender (Gender), the year the baby was born (Year), and the number of times that name was used that year (Count). Because the data is pulled from the SSA, the population includes citizens of the United States who have registered with the SSA who were born in years 1880 to 2018. The years before 1937 are not a complete picture of all names, since this is when the Social Security Administration was established, and not all people registered retroactively. Names that appear less than five times in a geographic population are not included in the data for privacy purposes [7].

Table 1 shows the format of the data from the Social Security Administration website [7]. It contains the birth year (Year), the name (Name), the gender (Gender) and the number of children born in that year given that name (Count). As is demonstrated here, 'Mary' is a popular girls name in the late 1800s, and is also uncommonly used as a name for male babies as well.

Table 1. Example Data from SSA for 'Mary', 1880-1884

Year	Name	Gender	Count
1880	Mary	F	7065
1880	Mary	M	27
1881	Mary	F	6919
1881	Mary	M	29
1882	Mary	F	8148
1881	Mary	M	30

By combining the data from 1918-2018, the top names in the last 100 years can be derived. Table 2 shows the the most popular male baby names and female baby names for years 1919-2018. The percentage describes what percentage of all children of that gender in the time period were given the name. For example, of all the boys born between 1918-2018, 2.82% of them were given the name 'James'.

Table 3 shows an example of the state level data provided by the SSA. The SSA provides the top five male and female names by state from 1980 to 2018. The data contains the state name (State) and Rank1-Rank5 baby names. Rank4 and Rank5 are not shown in the Table 3 example.

4.2 Context and Limitations of Social Security Administration Data

This organization was created in 1935 as part of New Deal legislation entitled the Social Security Act of 1935 [13]. Due to the social climate at the time, the

Table 2. Top Five Popular Names for Births, 1919-2018

	Male			Female		
Rank	Name	Count	Percent	Name	Count	Percent
1	James	4,722,254	2.82%	Mary	3,262,727	2.04%
2	Robert	4,494,870	2.67%	Patricia	1,560,583	0.98%
3	John	4,493,291	2.69%	Jennifer	1,467,196	0.92%
4	Michael	4,319,164	2.58%	Linda	1,447,912	0.91%
5	William	3,587,905	2.14%	Elizabeth	1,420,945	0.89%

Table 3. Example Data from SSA for Popular Girl Names by State, 1980

State	Rank1	Rank2	Rank3
Alabama	Jennifer	Amanda	Kimberly
Alaska	Jessica	Jennifer	Sarah
Arizona	Jennifer	Jessica	Melissa
Arkansas	Amanda	Jennifer	Melissa
California	Jennifer	Melissa	Jessica
Colorado	Jennifer	Jessica	Sarah

earliest years of this data are likely missing and misrepresenting several segments of the U.S. population, namely people of color and women. To understand why, a short history lesson about the SSA is needed.

Social Security as known now was established as a federal social insurance program. It required that workers contribute to the joint fund, thereby implying that only workers participating in the included industries were eligible to enroll. Only about 60% of the population in the United States was covered. There were some sectors that were excluded from participating, most notably of which were agriculture and domestic work. There is debate amongst historians about whether this was intentionally discriminatory against African Americans and women [4]. It took until 1950 for farm workers and domestic employees to be covered [17].

Regardless of the motivation, the fact of the exclusion means that the data set is likely quite biased, particularly for early years, to exclude black Americans and women. Evidence of this is shown in Table 2, where top boy name is 4.7 million, but the top girl name is 3.2 million. One way that this is accounted for is to give higher priority to insights found from more recent name trends. This allows for the changes in the inclusion of industries in the overall program, and also changes in how Social Security numbers are assigned. Beginning in 1987, the SSA established a program called “Enumeration at Birth”, which allowed parents to register their children as part of the birth registration process [12].

This means that data beginning after 1987 no longer requires that the registrant be working, and should show a better representation of name prevalence.

4.3 Exploring the Social Security Administration Data Visually

The SSA is deceptively simple data. The tables above give good examples of format, but further visual exploration of the data highlights additional richness.

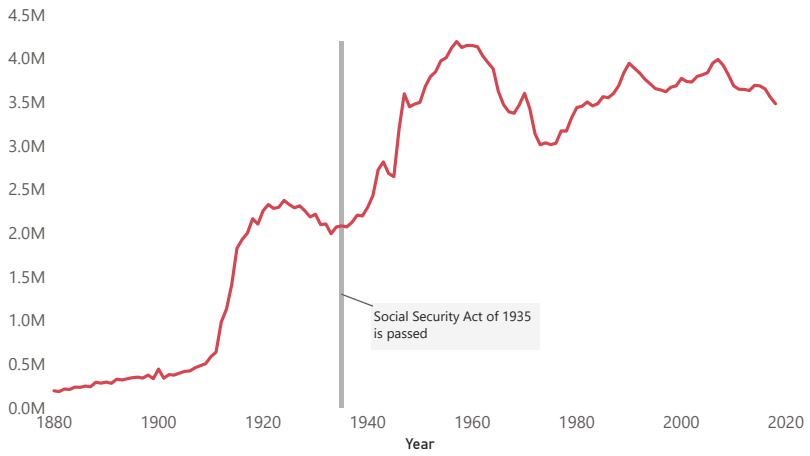


Fig. 1. SSA Registrants by Birth Year

Figure 1 demonstrates the changes over time in the registration of people with the SSA. Because of the policy changes that happened throughout the 20th century, there are variations in trend, with the last 30 years looking relatively stable in volume.

Looking more specifically into several names, it is evident that popular names do change over time, as demonstrated in Figure 2. Names tend to peak and drop off, and have different shapes. The height of the peaks varies by name, and the persistence of the name's popularity changes per name. One interesting feature to note on visual inspection is that these names all appear to have somewhat symmetric patterns. Some names, like 'Karen' (noted in dark blue in Figure 2), have a shallower sloped increase and decrease than others, like 'Jennifer' (in purple).

Many names can be considered derivations of other names. 'Jake' is often short for 'Jacob', and 'Mike' is short for 'Michael'. In general, the root name that the nicknames are derived from tends to have a more influential trend, as is expected from a governmental data set where full and proper names are typically used. There was an interesting shift from the name 'John' to a longer form, 'Jonathan', shown in Figure 3.

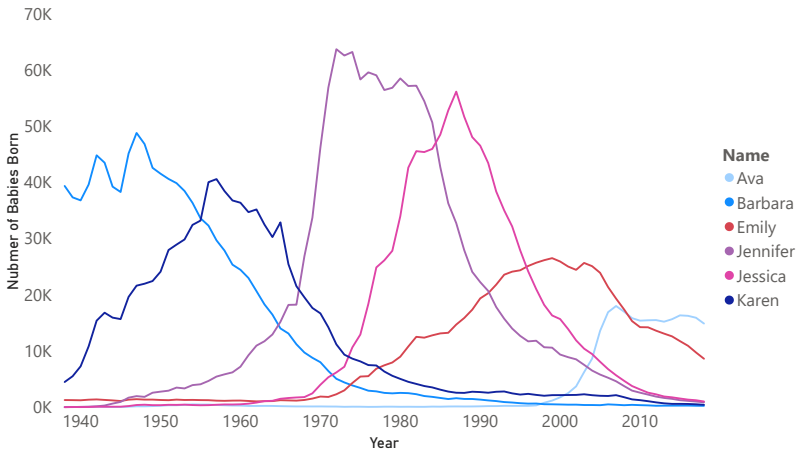


Fig. 2. Changes in Most Popular Girls Names

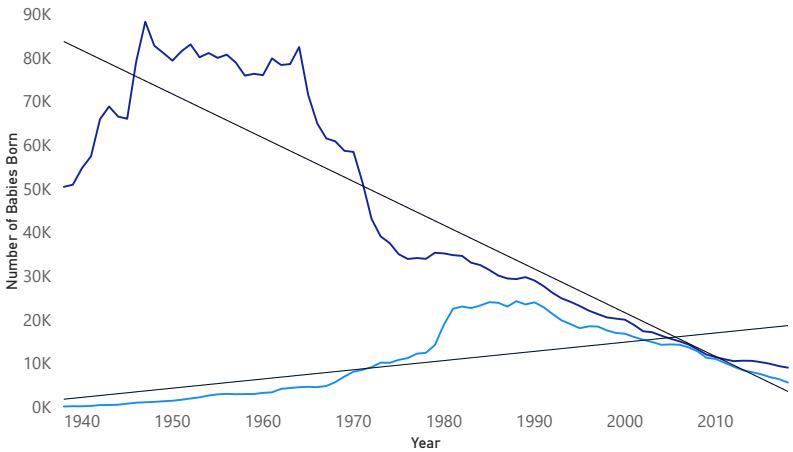


Fig. 3. Shift from 'John' to 'Jonathan'

Figure 4 highlights that there are similar behavior in names that have similar sounds. Figure 4a demonstrates the common decline of names that start with the letter 'D', and Figure 4b highlights a similar trend for names that end with the sound '-ary', such as 'Gary' and 'Larry'. The observed behavior of similar sounding names further motivates the experiment utilizing phonemes for prediction.

4.4 Historical Data

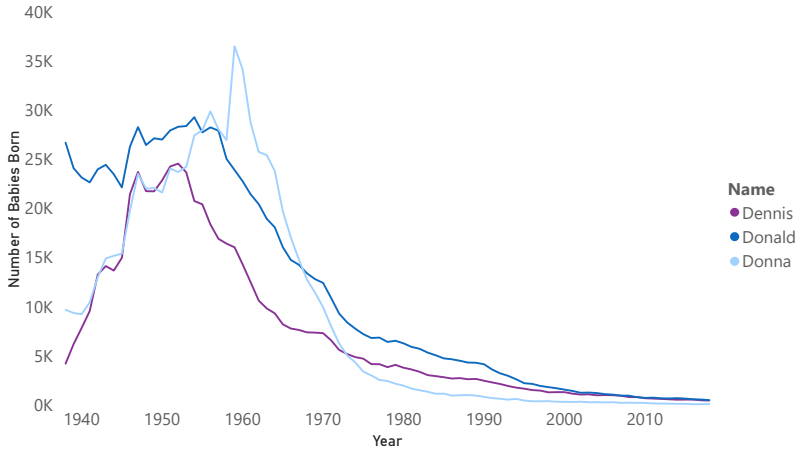
The historical names data set consists of the name of the influencer (the popular historical figure or hurricane), the peak year or years of popularity for the name, and the area of popularity for the name (author, actor, political figure, singer, hurricane). The peak year was chosen based on the year that the person would feature most prominently in the news, thus making their name top-of-mind for the general population. Table 4 indicates the defining moments used for each group of influencers. Any additional information that may be pertinent was also collected and stored in the historical names data set, such as the most popular movie that an actor starred in, the first best-selling book by an author, or the geographic area that a hurricane hit the hardest.

Table 4. Influencer Groups and Definitions

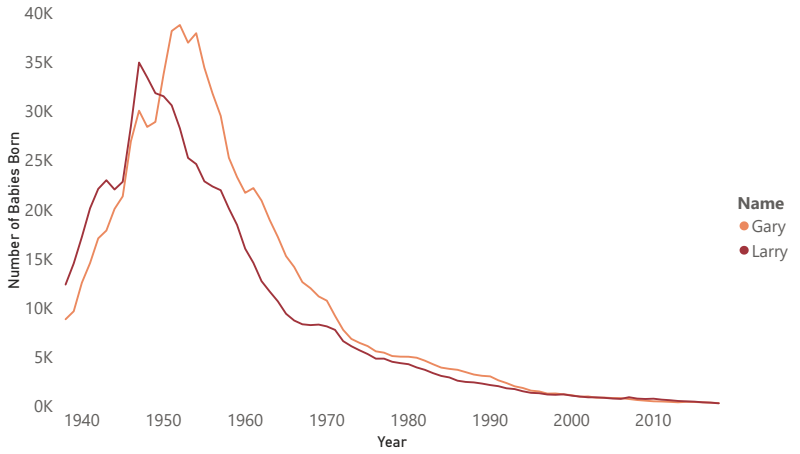
Group Name	Influencer Year Definition	Total Number
Actresses	First award-winning movie	100
Authors	First appearance on the best-seller list	272
Hurricanes	Name retirement	27
Athletes	First or most significant championship	10
Historical Figures	Rise to fame	47

In addition to looking at influencers that were happening in the present time, biblical names were evaluated for changes in trends that may reflect the changing national populace away from predominantly Christian-faith observers. The list of 558 biblical names represents approximately 19% of the total population of the United States in the 1918-2018 time period. Table 5 details how these names are distributed across genders.

Some of the best known names from the Bible are the writers of the Gospels: Matthew, Mark, Luke, and John. The shift in interest even among this small cohort of names can be seen in Figure 5a. 'John' and 'Mark' are popular names earlier on, with 'Matthew' peaking later, and 'Luke' not having an apparent peak in the time frame shown. It is also interesting to note the slight bump in the popularity of 'Luke' in the late 1970s and early 1980s, which corresponds to the first Star Wars trilogy release in which a lead character is named Luke. This is particularly evident when looking at the percentage change compared to the



(a) Names Beginning with 'D'



(b) Names Ending in '-ary'

Fig. 4. Similar Name Trend Behaviors

Table 5. Biblical Names

Gender	Number of Names
Male	424
Female	120
Male or Female	14

prior year in Figure 5b. The number of babies named Luke increased by over 50% from the previous year in 1978, which is the year after the first Star Wars movie was released.

4.5 Phoneme Data

To account for the various sounds in names, the Carnegie Mellon University Pronouncing Dictionary was utilized to build name phoneme vectors [14]. The Pronouncing Dictionary includes over 134,000 words in the English language. Each entry includes the word and the phonemes that compose that word, examples of which can be seen in Table 6. The grouping of letters correspond to the phonetic in the word. The numeric characters in the source indicate stress values used to correctly emphasize syllables in the word. For this paper's purpose, these numeric values were stripped from the phonemes before being vectorized for use in the neural network.

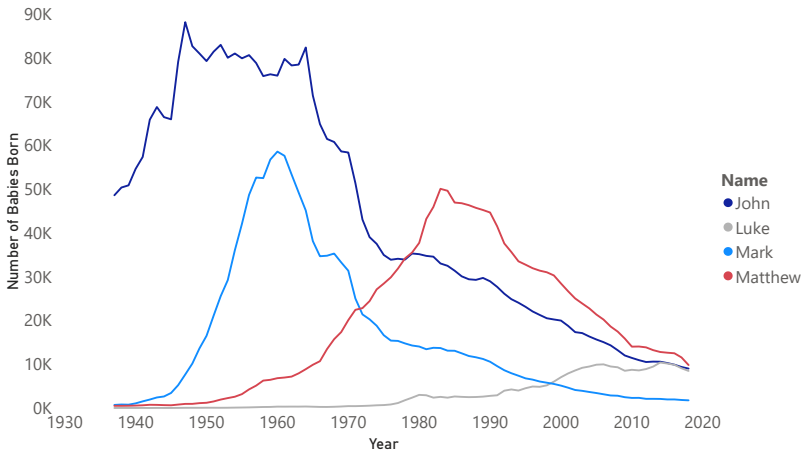
Table 6. Pronouncing Dictionary Entry Examples

Word	Source Phonemes	Project Phonemes
EXAMPLE	IH0 G Z AE1 M P AH0 L	IH G Z AE M P AH L
WOMEN	W IH1 M AH0 N	W IH M AH N
SUCCESS	S AH0 K S EH1 S	S AH K S EH S

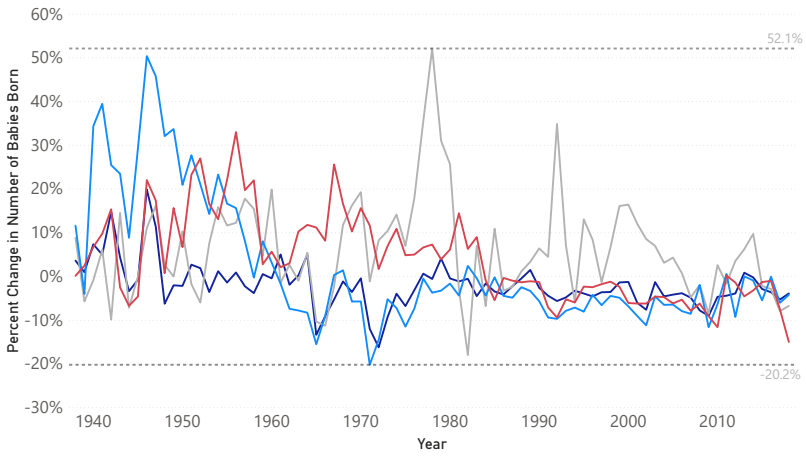
The examples given are of non-name words, but the Pronouncing Dictionary includes many names. In order to extract them, a comparison was made between the names from the Social Security Administration data and the words in the dictionary, and the matching phoneme sets were extacted. Not all names appeared in the dictionary, but 15,704 names with corresponding phonemes were gathered from it. For the purpose of exploration, this was considered a large enough example to demonstrate the success of phonemes in predictive power. It represents approximately 1.6% of the overall name population in the Social Security Administration data, but also includes the majority of the most popular names.

5 Influencer Index

To determine if an person from the influencer data sets had an impact on the popularity of their name, an index function was utilized. This function was originally defined in "Naming Boys after U.S. Presidents in 20th Century" [9]. One important consideration to account for was that influential events may happen anytime during the year, meaning the impact may be be seen in the same or the year following an influential event. The year of influence and the



(a) Gospel Name Popularity



(b) Percent Change in Popularity from Previous Year

Fig. 5. Name Trend of Gospel Writers' Names

year following was examined and the maximum name count was taken as the basis for the index calculation.

This maximum value was the normalized by using the median and interquartile range of the five years prior to the year of influence. Equation 1 shows how these components relate to one another. The effect of doing this normalization is to remove the effects of an existing trend and highlight the points in time where changes are significant in a given year.

$$c = \frac{y - \widehat{q}_{0.5}}{\widehat{q}_{0.75} - \widehat{q}_{0.25}} \tag{1}$$

A look-forward method was used, meaning that the analysis started with the curated list of influencers to find impacts from those people in their year of greatest influence. This was the most systematic way of gathering derivable groups of people, that could then be evaluated in real-time to feed the prediction model. In order to be considered a valid influencer, the magnitude of the index score needed to be greater than 2. This means that the year of influence count was at least two times as great as the normal count in the previous five years.

6 Phoneme Experiment

The models used in this analysis were built with Tensorflow and ran on the Google Colab GPU for faster processing. Even so, computational constraints were encountered for the phoneme-driven model. Both models are Long Short-Term Memory (LSTM) neural networks. The LSTMs consist of three layers: embedding, lstm, and a single dense layer. There is also an input layer that sits on top of the embedding layer. The single dense layer at the end is a binary predictor: if the name count increases, the value was 1, and 0 otherwise. This is necessary to give the LSTM a target for which to train and a way to measure the accuracy of the models. The structure of the LSTM is shown in Figure 6 below.

The LSTM layer has 100 neurons and the dense layer implements a sigmoid activation function, as it is predicting a binary output. The models implement binary cross entropy as the loss function and the adam optimizer to optimize the loss function. To evaluate the model, the accuracy metric was used. The models were trained for 20 epochs with a batch size of 32.

This experiment consisted of using identical structures for both models. The first model was given the names based on name count along with 68 years of popularity data as inputs, 1951-2018. The second model was given vectorized phonemes and the last thirteen years of count data. This choice was made for performance and tuning of the LSTM parameters, as the second LSTM required significant additional computational work for estimating the phoneme embeddings. The input difference between the models allows us to consider that other variables are held relatively equal, to allow for attribution to the phoneme construct and not solely the name count.

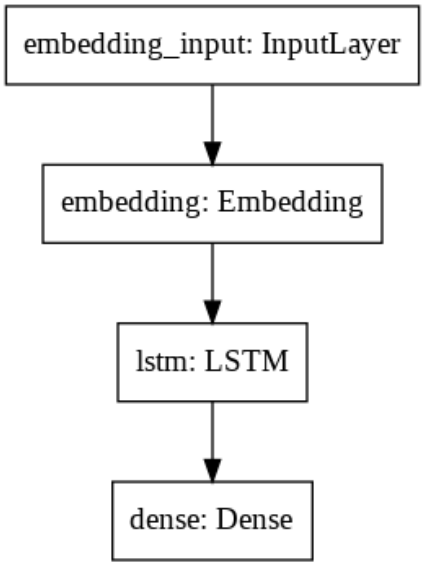


Fig. 6. LSTM Layers

7 Results

7.1 Influencer Index Results

The indexing score was applied to each of the historic names. Figure 7 is an example of the name 'Diana'. In this graph, the popularity of the name Diana has a large peak in the 1950s and 1960s, and has a resurgence in the early 1980s. Lady Diana began dating Prince Charles in 1980, and the spike in her name suggests that she may have influenced the name trend. Figure 8 shows the what the influence score would be for each year, if there was someone who was influential with that name in that year. There is a very large spike in the influence score for 'Diana' in 1980, which further suggests that she was a contributing factor.

The majority of the selected names did not generate a high enough score to be considered an influence on the name trend. Figure 9 shows the overall percentage of people in each group that had a score over 2. None of these groups appears to have a strong predictive capability when it comes to name trends.

7.2 Historical Figures Results: Biblical Name Trends

Approximately 19% of the naming data from the SSA represents names that occur in the Bible. The heatmap in Figure 10 represents the percentage change from the previous year for each biblical name. Blue color indicates a decrease in name count from the prior year, with darkest blue hues indicating greatest

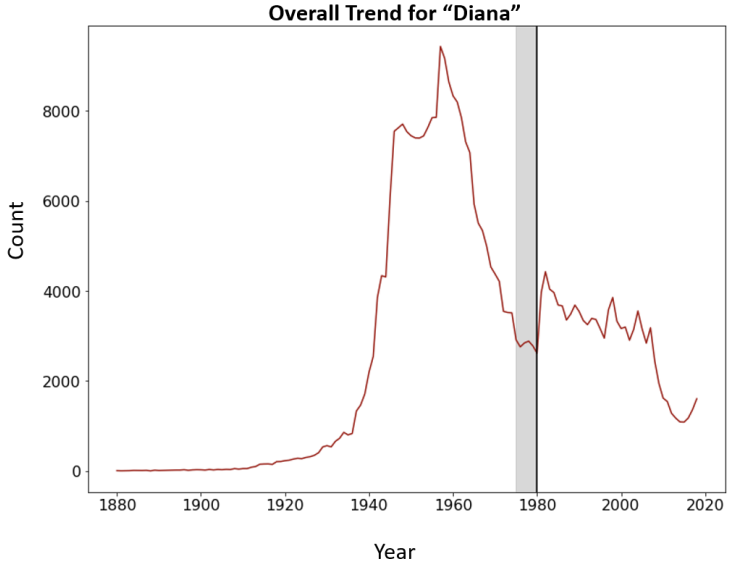


Fig. 7. ‘Diana’ Name Trend with Lady Diana’s Influence Year

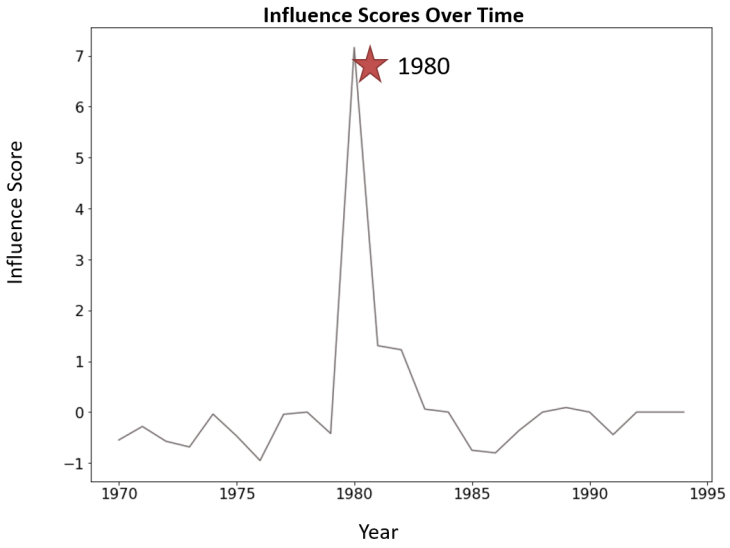


Fig. 8. Influence Score for ‘Diana’ for 1970-1995

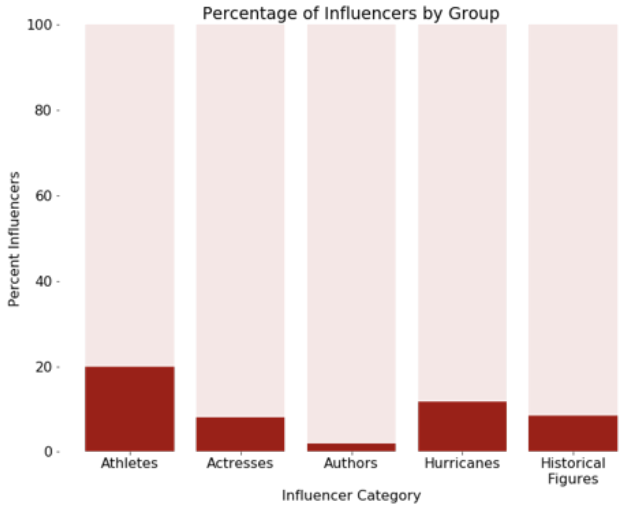


Fig. 9. Influencer Index Success

decreases. Red color indicates an increase in the name count from the prior year, with the darkest red hues indicating greatest increases. The actual values in the graph are not important. The key takeaway is that biblical names are more steadily declining, based on the highly concentrated blue values on the right side of the timeline. This subset of names represents the biblical names that also appear in the top 200 names from 1937 to 2018, and the same pattern appears for less popular biblical names as well.

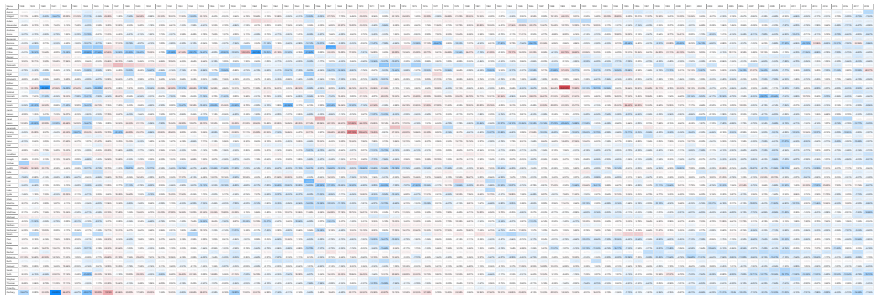


Fig. 10. Heatmap of Percent Changes in Names Over Time[4]

[4] A high-resolution version of this figure is available at https://github.com/laurajludwig/Analyzing-Influences-on-US-Baby-Name-Trends/blob/master/heatmap.pdf.

7.3 Phoneme Model Results

Table 7 shows the side by side comparison of the two model results: Model1 being the count-driven model and Model2 being the phoneme model. The accuracy metric for the models reflects the fact that the phoneme model was correctly predicting the trend to go up or down about 72% of the time, which is not poor performance when compared to the baseline model with 92% accuracy, which was trained on 55 more years of data. It should be noted that this analysis only examined the accuracy metric, which can be misleading. Further work could include evaluating the models with the F1 score as well, which penalizes the incorrect classification that a name wasn't going to increase but did increase. However, because the goal of the phoneme-driven model was to extract interesting name embeddings, there was no concern for the specific evaluation metric other than for comparison with the count-driven baseline model.

Table 7. LSTM Model Results and Summary

Model	Accuracy (%)	Years of Data
Model1	91.81	1951-2018
Model2	72.28	2006-2018

The name embeddings that were generated by the second LSTM model based on the phonemes were extracted and examined. Figure 11 shows some names that are deemed similar according to the embeddings. These groupings seem to have some similar sounds. For example, 'Ashley' has an 'l' in a similar position to 'Delima' and 'Cal'.

```
{'ASHLEY': ['ELMORE', 'CARYN', 'DELIMA', 'BIRK', 'CAL'],
 'BILL': ['LUE', 'CLEOPATRA', 'GUNNAR', 'CHARISMA', 'ALYS'],
 'JESSICA': ['KENN', 'CRAWFORD', 'JUMANJI', 'MARIO', 'PRAIRIE']}
```

Fig. 11. Similar Name Embeddings Generated by Model2

The embedding layer consisted of vectors that contained 80 variables. The most important components of these vectors were projected into two-dimensional space using t-distributed stochastic neighbor embedding (TSNE). This is a method of feature-space reduction that allows for the visualization of a large number of variables in just a few dimensions. Clusters that formed were examined as they may be indicative of a phoneme similarity feature that could be used to predict the next name. Figure 12 shows one example of a cluster of names. Not all name embeddings were visualized in clusters, only a random subset.

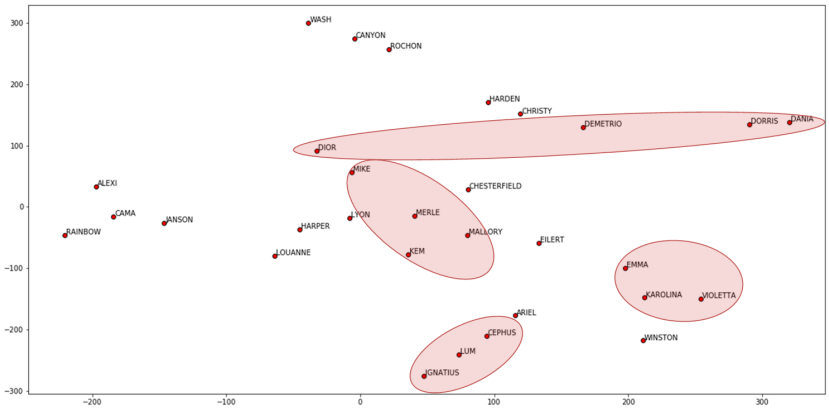


Fig. 12. Clustered Similar Name Embeddings, Model2

8 Analysis

8.1 Influencer Index Analysis

The low percentages of influencers in each of the selected groups from the influencers analysis suggests that selecting people based on their most influential year is not the correct approach to use in identifying influencers. The influencer index calculation worked for some influential people, but not for others. It is likely that there is another feature of these influential individuals that was not accounted for here. Working backwards from high influencer scores to define key name-year pairs and then searching for those influencers was somewhat successful in ad-hoc evaluation (example: the 'Luke' observation about Star Wars).

8.2 Phoneme Experiment Analysis

The accuracy score for the phoneme-driven model indicates that the pronunciation of names contributes to prediction, but is not necessarily the driving force behind name trend. Reviewing the name embeddings generated by the LSTM indicates that the model is able to identify similar sounding names based on the clustering tendencies, as shown in Figure 12. The model clusters names that start with a 'D' on the same x-axis line: 'Dior', 'Demetrio', 'Dorris', and 'Dania'. However, 'Dior' is clustered more closely to 'Mike', possibly because of the strong 'I' sound at the beginning of both names. Female names that end in the similar, soft 'A' sound consist of their own cluster with the names 'Emma', 'Karolina', and 'Violetta'. The fact that 'Dania' starts with a 'D' must be more significant to the model for prediction than the ending sound of the name, since it is clustered with the 'D' names instead of the soft sounding 'A' end names. The names that start or contain an 'M' are also loosely clustered: 'Mike', 'Merle', 'Mallory', and 'Kem'. The names that contain or end in a similar 'U' sound are similarly clustered: 'Ignatius', 'Lum', and 'Cephus'.

However, the visualization of these names provides several strange results as well. 'Chesterfield', a very rare name, is clustered close to 'Mallory' and 'Eilert' (also a rare name). None of these names have similar pronunciations, but perhaps they have similar trends or are all similarly rare. The two names that start with 'W', 'Wash' and 'Winston', are not clustered together, and neither are the names that start with 'H': 'Harden' and 'Harper'. There are additional name clusters that are not intuitive, such as 'Rainbow' and 'Cama'. This clustering is perhaps due to the second letter 'A' in both names, but otherwise the names do not have similar pronunciations. These results could possibly explain the lower accuracy metric for the phoneme-driven model. Longer training and more data may allow the model to correctly cluster all these names based on their pronunciation. But as is, the model impressively captures multiple dimensions of name pronunciation.

9 Ethics

Two ethical considerations within this analysis are around privacy and bias. The aggregation of data protects individuals from being identifiable directly in the data set. The Social Security Administration restricts this to some extent by only providing the data for which there are at least 5 people in a geographic area born in the year with that given name. This applies at both the state and national level. If there are 5 people with a unique name in all different states, their name will appear in the national data, but will not appear in the state level data for that year. This guards individuals from being directly identified, but it also limits the scope of the analysis that can be done. This trade off is necessary from the data owner's perspective in order to preserve privacy for individuals.

The second ethical consideration is bias. As mentioned in Section 4 about data, there are some fundamental limitations in how this data was collected, namely that it excluded important populations. The conclusions from this analysis and from other analysis using this data must be limited to account for the social biases during the earlier years of the available data. The attention was focused on the last 30-40 years of data in order to limit the effects of the earlier social climate, but the early data was not outright excluded. The focus of the phoneme model on the last thirteen years, in part for performance, but also to account for the rapidly changing social culture, in order to draw relevant and unbiased conclusions about the usefulness of phonemes in predicting baby names.

10 Conclusions

These findings suggest that identifying influencers based on a common milestone is not sufficient for predicting what names will become popular next. There is evidence to support the statement that cultural icons and other newsworthy names influence name trends, but not in a systematic useful way that can be modeled consistently. The only cohort of names that exhibited consistency in

behavior was biblical names, all of which were decreasing in popularity over the last 20-30 years. Predictions of biblical names can be penalized in a model accordingly.

The results of the phoneme-based model suggest that similar sounding names provide an alternate, if not ideal, method for predicting if a name will increase in popularity. While the phoneme-driven model did not perform as well as the baseline model, training the model on more data, with more epochs, and with hyperparameter optimization, would likely increase performance. However, transforming phoneme sets into vectors that are processed in a similar fashion to words was a productive experiment as visualizing the results provided insight into how the LSTM predicts name trend. In order to optimize performance and minimize loss, the model learned similar embedding vectors for similar sounding names, indicating a relationship between name trend and name pronunciation.

References

1. Brownlee, J.: What Are Word Embeddings for Text? Machine Learning Mastery (2017), https://machinelearningmastery.com/what-are-word-embeddings/, last Updated: August 2019
2. Burdess, N.: Whats in a name? a brief history of baby name trends from the anglo-saxons to today. Online (October 2017), https://www.historyextra.com/period/norman/baby-names-popular-royal-history/
3. Claudia Deane, M.D., Morin, R.: Americans name the 10 most significant historic events of their lifetimes. Online (December 2016), https://www.people-press.org/2016/12/15/americans-name-the-10-most-significant-historic-events-of-their-lifetimes/
4. DeWitt, L.: The decision to exclude agricultural and domestic workers from the 1935 social security act. Social Security Bulletin **70** (2010)
5. Francis, D.R.: Employers' replies to racial names. National Bureau of Economic Research (September 2003)
6. Fryer, Roland G., J., Levitt, S.D.: The Causes and Consequences of Distinctively Black Names*. The Quarterly Journal of Economics **119**(3), 767–805 (August 2004). https://doi.org/10.1162/0033553041502180, https://doi.org/10.1162/0033553041502180
7. Government, U.S.: Social security. This website is produced and published at U.S. taxpayer expense., https://www.ssa.gov/
8. Jonah Berger, Eric T. Bradlow, A.B., Zhang, Y.: From karen to katie: Using baby names to understand cultural evolution. Association For Psychological Science **unknown** (2012), to appear
9. K. Kuakowskia, P. Kulczyckia, K.M.A.D.P.G., Krawczyka, M.: Naming boys after u.s. presidents in 20th century. Acta Physica Polonica A **129** (2016), to appear
10. Levitt, S.D., Dubner, S.J.: Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. Harper Collins, 195 Broadway, New York, NY 10007, 1 edn. (September 2011)
11. Nicholson, C.: A Beginner's Guide to Neural Networks and Deep Learning. Skymind, https://skymind.ai/wiki/neural-network

12. Puckett, C.: The story of the social security number. Social Security Bulletin **69**(2) (2009), https://www.ssa.gov/policy/docs/ssb/v69n2/v69n2p55.html
13. of Representatives, U.H.: The social security act of 1935 (August 1935)
14. Rudnicky, A.: The CMU Pronouncing Dictionary (2014), http://www.speech.cs.cmu.edu/cgi-bin/cmudict, version cmudict-0.7b
15. Tolga Bolukbasi1, Kai-Wei Chang, J.Z.V.S.A.K.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. arXiv **unknown** (2016), to appear
16. Vincentelli, E.: You are what your name says you are. The New York Times pp. 43–43 (November 2007)
17. William J. Nelson, J.: Employment covered under the social security program, 1935-84. Social Security Bulletin **48**(4) (April 1985), https://www.ssa.gov/policy/docs/ssb/v48n4/v48n4p33.pdf
18. Zhongheng Zhang, Marcus W. Beck, D.A.W.B.H.W.S., Goyal, H.: Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Annals of Translational Medicine **6**(11) (June 2018)